语言学知识驱动的空间语义理解能力评测数据集研究*

詹卫东，孙春晖，肖力铭

（北京大学　中文系　北京　100871）

提　要　近 20 年来，深度学习技术显著提升了机器的自然语言处理能力，使之在诸多任务上接近甚至超过人类水平。机器学习的对象不再是直接来自人类语言学研究成果（知识），而是人类语言材料（数据）。在靠数据和算力驱动的大语言模型几近建成巴别塔的当下，语言学家通过深挖语言现象总结的语言学知识价值何在？本文提出从知识到数据的研究思路，设计了空间语义理解的 6 项任务：空间信息正误判别、异常空间信息识别、缺失参照成分补回、空间语义角色标注、空间表达异形同义判别、空间方位关系推理，以构建中文空间语义理解能力评测数据集为例，介绍从 SpaCE2021 到 SpaCE2024 数据集的设计思想、数据集制作概况以及机器在空间语义理解任务上的表现。总的来看，参加 SpaCE 赛事的大语言模型，在依赖表面分布特征（形式线索）的任务上容易获得好成绩，在依赖深层语义理解（认知能力）的任务上容易表现不好。因此，在人工智能高速发展使得语言学知识在计算机信息处理领域被动边缘化的当下，语言学知识的价值需要拓展，即用于指导小而精的高品质语言数据，以提升机器学习的效果和效率。为了计算应用的目的，语法研究应该在观察充分、描写充分、解释充分之上，追求更具挑战性的目标——生成充分。

关键词　人工智能；大语言模型；语言学知识；空间语义理解；数据合成

中图分类号　H002　**文献标识码**　A　**文章编号**　2096-1014（2024）05-0007-15

DOI　10.19689/j.cnki.cn10-1361/h.20240501

SpaCE: A Linguistic Knowledge-Driven Benchmark for Spatial Cognition Evaluation

Zhan Weidong, Sun Chunhui and Xiao Liming

Abstract　Over the past two decades, deep learning technology has propelled machine natural language processing capabilities to rival or even surpass human levels in many tasks. Machine learning does not directly utilize the outcomes of human linguistic research (knowledge), but rather from human language materials (data). This situation should garner significant attention from linguists. As large language models, driven purely by data and computational power, have nearly constructed a modern Tower of Babel, the question of how to realize the value of linguistic knowledge through in-depth exploration of specific and subtle language phenomena looms large over every linguistic researcher. This paper proposes a research approach that generates text data from linguistic knowledge for evaluating machine understanding of spatial semantics. Over the past four years, we have

*　作者简介：詹卫东，男，北京大学教授，主要研究方向为计算语言学、语言知识工程、中文信息处理。电子邮箱：zwd@pku.edu.cn。孙春晖，男，北京大学在读博士研究生，主要研究方向为语言知识工程、自然语言处理。电子邮箱：psysunch@163.com。肖力铭，男，北京大学在读博士研究生，主要研究方向为语言知识工程、自然语言处理。电子邮箱：lmxiao@stu.pku.edu.cn。

教育部人文社会科学重点研究基地重大项目“面向机器语言能力评测的综合型语言知识库研究”（22JJD740004）。本研究得到北京大学计算语言所穗志方教授和常宝宝副教授课题组、复旦大学计算机学院邱锡鹏教授课题组大力支持；北京大学中文系多位同学参与了本文的工作。SpaCE2021、SpaCE2022、SpaCE2024 评测大赛由华为公司提供奖金资助。特此致谢。

organized four consecutive competitions on Chinese Spatial Cognition Evaluation (SpaCE): from SpaCE2021 to SpaCE2024, including 6 sub-tasks: Determination of Spatial Information Validity, Detection of Spatial Anomalies, Recovery of Spatial References, Identification of Spatial Semantic Roles, Recognition of Spatial Equivalences, and Spatial Position Reasoning. This paper introduces the design philosophy, dataset creation process, dataset overview, and the performance characteristics of machines in SpaCE tasks. Overall, large language models participating in the SpaCE competitions perform relatively well on tasks that rely on surface distribution features, that is, tasks with formal cues, but poorly on tasks that depend on deep semantic understanding, that is, tasks requiring cognitive abilities. In the current era of rapid AI development, where linguistic knowledge is passively marginalized in the field of natural language processing, the value of linguistic knowledge needs to be redefined. It should be used to guide the production of small, high-quality language data to enhance the effectiveness and efficiency of machine learning. For computational applications, grammatical research should pursue more challenging goals—adequate generation—beyond the objectives of adequate observation, description, and explanation.

Keywords artificial intelligence; large language models; linguistic knowledge; spatial semantic understanding; data synthesis

一、引　言

ChatGPT等大语言模型的问世，引发了对以乔姆斯基理论为代表的语言学研究理念的尖刻批评，如Piantadosi（2023），Hinton（2024）。语言学家不甘示弱，做了针锋相对的回击，如Katzir（2023），Chomsky，Roberts & Watumull（2023）。

乔姆斯基在《语言知识：性质、来源及使用》（Chomsky 1986）一书中提出了两个发人深思的问题。一曰“柏拉图问题”（Plato’s problem）：为何人能在证据严重不足的条件下知道如此之多？二曰“奥威尔问题”（Orwell’s problem）：为何人在证据充足的情况下却又如此无知？前者关乎个人认知，是乔姆斯基提出的“语言先天论”的主要依据[①]；后者关乎社会认知，是乔姆斯基政治评论的核心关切[②]。这两个问题有着强烈的“冲突张力”。其中所谓的“证据”（evidence），可以理解为一般常说的“数据”（data）；而与“数据”相对的，则是人所知道的“知识”（knowledge）。这样来看，乔姆斯基提出的这两个问题，实际上共同关联着一个更为基本的问题，即“知识”和“数据”之间到底是怎样的关系。更进一步，在今天大语言模型引领的人工智能（AI）时代，面对以ChatGPT为代表的大语言模型有时表现出的堪比人类水平的自然语言生成和理解能力，作为语言学研究者，很自然地会沿袭乔姆斯基的提问方式，生发出这样的疑问：为何机器能在不需要语言学知识加持的条件下，获得如此惊人的语言能力？为何人在语言学知识如此丰富的条件下，却始终未找到帮助机器把语言学知识转化为语言能力的可行途径？

本文并不打算展开探讨上面这两个宏大的问题，而是尝试以这两个问题为背景，从机器学习的视角，重新思考如何认识知识与数据之间的关系。我们相信，要在AI时代更好地发展语言学研究，发挥语言学研究成果的价值，更需要坚持实证主义的研究路径，通过大量不断地与机器的语言交互，来深入考察和分析机器在自然语言相关任务上的“能”与“不能”，从而深化对机器以深度学习方法学

① 大意是：儿童在缺少言语刺激条件下很短时间内就可以掌握语言（即拥有丰富的语言知识），只能归结为语言知识本就是人脑（先天）官能的一部分。

② 大意是：现代社会中权力结构会无孔不入，制造“共识”并操控“共识”，从而导致“罪行证据比比皆是，人们却一无所知”（乔姆斯基的学生、哈佛大学黄正德教授语）。

习人类语言这种特殊方式的理解，同时也加深我们对人类自身语言能力和语言学知识之间关系的理解（詹卫东 2024）。

需要强调的是，**数据**是具体可观察的，在本文中特指**语言材料**；**知识**是抽象的模型，在本文中如不说明则特指**语言知识**。语言学家提出的各种语言学理论，称为“**语言学知识**”，是人类对于语言知识的想象和外化，而非语言知识本身。

二、机器空间语义理解能力评测任务设计

人与机器的关系，可类比人类语言教学中教师与学生之间的关系，可以分为教学和测试两个方面来看。

先看教学，在人类尝试让机器具备人类自然语言能力的早期探索阶段，是完全按照人类师生教学的模式进行的，即把人总结出的语言学知识（词典和语法规则）转换为形式语言表述的结构化知识库，作为语言模型教给机器。构建这样的语言模型，主要依赖人的洞察力，或者直接由人工发掘，或者借助计算机辅助人来发掘，都是以显式符号为基础、对人而言可理解的知识表征。这就是所谓“符号主义”（Symbolism）的人工智能研究路径。后期崛起的“联结主义”（Connectionism）人工智能则完全是另一条路径，机器学习人类语言的方式从向人学习逐渐发展演变成了以“自学”（Self-Supervised Learning）为主。受人类大脑的神经网络工作原理启发，计算机科学家设计了多层深度人工神经网络，直接通过“输入字符串–输出字符串”的数据配对（可大致理解为“问题–答案”样例），学习一个能够映射“输入–输出”数据的函数，即所谓“端到端”（end-to-end）的学习方法，使得机器在给定输入字符串条件下，能得到正确的输出字符串结果（参见 Wolfram 2023）。以“张三是县长 ___ 来的”为例，如果以这个缺失了词语的句子作为输入，输出字符串可以是“派 / 请 / 抓 / 昨天招聘来”等等。把这样的“输入–输出”数据对，喂给深度神经网络学习，在句子数量足够多、神经网络参数足够大的情况下，机器最终就可以捕捉到汉语中任意词语之间在特定语境条件下的依赖关系，从而表现出能够理解句子意思以及生成出自然句子的能力。

再说测试。机器的语言能力和智能水平，需要通过测试来检验（Legg et al. 2007；Chollet 2019；董青秀，等 2021）。针对机器的测试，大致有 4 类做法。（1）可看作考语言学知识，比如让机器完成中文分词、词性标注、句法结构分析等任务，机器需要掌握词、词类、句法结构层次分析等语言学专业知识。但这种测试方式在大语言模型出现后基本已经行不通了，因为语言学知识和语言实际应用能力之间并无必然的关系，从最终应用的角度看，人们希望机器具备实用的**语言能力**（比如翻译、写文章等），而不是具备**语言学能力**。（2）直接拿考人的题目来考机器，比如用高考这样的标准化考试来考大语言模型。可参看 Zhong et al.（2023）。（3）大型综合性评测。具体又可分为两种不同的情况：一种是有较为系统的测试体系，并且由程序来判分，如 SuperCLUE、C-Eval、OpenCompass 等大型测试平台[①]；另一种是不设测试体系，由人类投票，采用 Elo 等级打分系统来评分，如大模型盲测竞技场 LMSYS Chatbot Arena[②]。（4）专项考试。这种方式相当于单科测试，一般聚焦于考察机器某一特定方面

① SuperCLUE 网址：https://www.cluebenchmarks.com/；C-Eval 网址：https://cevalbenchmark.com/；OpenCompass 网址：https://opencompass.org.cn/。

② LMSYS 网址：https://chat.lmsys.org/。关于 Elo，可参考 https://zh.wikipedia.org/wiki/ 等级分。

的能力，比如考察常识推理能力的 Winograd 挑战赛及其升级版 Winogrande 挑战赛[①]。

本文介绍的SpaCE评测研究工作[②]，属于上述第四类专项考试的范畴，考试科目可以概括为中文空间信息语义理解。开展这项研究的动机是探索以语言学知识来指导具体测试任务的设计和数据集的制作。概括来说，我们的指导思想有二。（1）从“形式—意义”对应关系的视角看机器的语言能力，区分“形式—意义”配对容易和困难的问题。之所以选择空间信息语义理解这个主题，是因为空间信息的主要语言标记方位词属于指示语（deixis）范畴。跟实词（如“国王、女人、旅行”等）不同，指示语的具体意义，需要依赖上下文和现实世界的情境，其“形式—意义”的对应关系超越了符号字面形式。人在理解的时候，需要调用更深的认知加工能力，才能在符号跟现实世界之间建立正确的联系。（2）测试任务应有层次性和结构性，应能从多个维度和不同深度探测机器的语言能力。探测结果不是一个简单的分数，而应该是机器语言能力的一个细粒度的综合呈现，类似于一个详细的体检报告。第一点认识，是我们选择空间领域作为测试主题的原因；第二点认识，则是我们进一步剖析空间领域内的具体问题，规划测试子任务的工作依据。

下面逐一介绍针对文本空间信息理解设计的6个任务，大体上遵循语言学中“语法—语义—语用”的递进关系来展开。在 SpaCE2021 到 SpaCE2024 的评测赛事实践中，后一届赛事相对于前一届，基本上是一个不断增加任务类型的过程，SpaCE2024 覆盖了这 6 项任务中的 5 个，是任务类型最多的（参见下文表 8）。

（一）空间信息正误判别

请看下面两个例句。其中例（1a）是从自然语料中抽取的富含空间信息的一个段落，例（1b）是把（1a）中“遂右转弯由东向西行驶”替换成了“遂右转弯由西向东行驶”。

（1）a. 大客车沿新源路由北向南行驶至曹安公路路口处遇绿灯，遂右转弯**由东向西**行驶，适逢被害人李红英骑电动自行车沿新源路西侧非机动车道由北向南行驶至此，两车相撞。

b. ……遂右转弯**由西向东**行驶，……

图 1 显示了例（1b）中的空间信息冲突。大客车由北向南行驶到十字路口右转弯，其行进方向只能是由东向西，而不是由西向东。（1b）文本中蕴含的这个空间信息冲突，涉及对同一个实体（大客车）的 3 个空间信息描述“由北向南”“右转弯”“由西向东”在时序上无法衔接这一空间常识知识的理解。

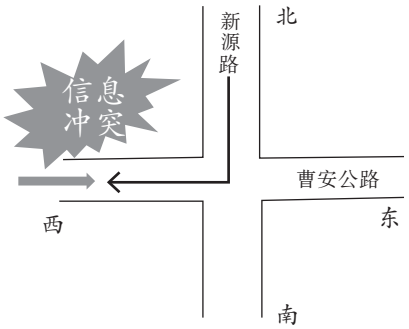

图 1 “右转弯由西向东行驶”空间信息冲突示意图

当把例（1a）和（1b）这样的句对呈现给计算机的时候，计算机应该能像人一样，判断（1a）中的空间信息是正确的，符合常理；（1b）中的空间信息是错误的，与常识相悖。

（二）异常空间信息识别

在判断一段话中存在异常空间信息的同时，实际上也应该能清晰地将异常信息的片段抽取出来，这就是比文本空间信息正误判断更具体的文本中异常空间信息识别任务。下面来看一个文本中包含异常空间信息的例子。

① 关于 Winograd 挑战赛，详见 Levesque et al.（2012）；关于 Winogrande 挑战赛，详见 Sakaguchi et al.（2020）。

② SpaCE 是 **Spa**tial **C**ognition **E**valuation 的缩写。近 4 年我们依托中国计算语言学大会（CCL）的中文技术评测平台，组织了 SpaCE 系列评测大赛 SpaCE2021 ～ 2024。网址：https://2030nlp.github.io/SpaCE2024/。

（2）夫妻俩商量了几天，买了一辆农用三轮车。农用三轮车的油门在左把手上，张顺东请人改到左边，他就能用左手操作了。

例（2）中的异常空间信息在于“农用三轮车的油门在左把手上，张顺东请人改到左边”这个片段。既然已经“在左把手上”，还要“改到左边”，就不合常理，形成信息冲突。如果用自然语言来描述（2）中的异常空间信息，不同的人可能会有不同的表达方式。为便于对机器的答案进行自动评分，最好是对标注格式进行统一规范。为此，我们提出了 S–P–E 空间三要素标注法，其中 S 代表空间实体，P 代表空间方位信息，E 代表跟 S–P 有关的事件信息（一般 E 由动词表达）。表 1 展示了对例（2）的 3 种 S–P–E 标注形式。很显然，第 1 种标注各要素对应整齐，最为合理；另外两种标注，P 要素或 E 要素信息标注各有不合理的地方。限于篇幅，这里不展开讨论，有兴趣的读者可访问 SpaCE 网站查询关于 S–P–E 标注的详细规范。[①]

表 1　S–P–E 三要素标注法标注文本异常空间信息示例

序号	异常片段 1			异常片段 2			异常类型
	S1	P1	E1	S2	P2	E2	
1	油门	左把手上	在	油门	到左边	改	语义冲突
2	油门	在左把手上	—	油门	—	改到左边	语义冲突
3	油门	在左把手上	—	油门	到左边	改	语义冲突

（三）缺失参照成分补回

中文常见的表达空间方位的形式是“名词 + 方位词”，如“教室里面”“桌子上面”。但行文中方位词前的名词常常也会出现承前省略的现象。例如：

（3）a. 这 20 管试剂都被封存在一个长方形的纸箱里，上面贴了一张白色的标签。

b. 这 20 管试剂被封存在一个长方形的透明玻璃箱里，上面事先都贴了不同颜色的标签。

例（3a）和（3b）逗号后面小句开头的“上面”这个方位词不是紧跟在名词之后，这个“上面”依赖前文哪个名词，需要联系上下文，在理解句子中实体之间语义关系的基础上才能确定。这两个句子的词语差异并不是很大，但“上面”所依赖的参照成分却是明显不同的：（3a）的空间信息应解读为“**纸箱上面**贴了一张白色的标签”，（3b）则应解读为“**20 管试剂上面**都贴了不同颜色的标签”。人有能力准确地理解整句所表达的空间场景信息，其中就包含一种能力，即在方位词前面补出缺失的空间参照成分。

（四）空间语义角色标注

从语言学角度讲，人对文本中空间信息的理解能力不仅可以通过前面 3 个任务体现，还可以进一步通过对文本中空间信息的结构化分析，在更深的层面上做更细粒度地刻画。下面例（4）是前文例（1）的更完整的文本。表 2 就是对这一文本中空间信息的结构化标注。

（4）2020 年 7 月 16 日 7 时 8 分许，牌号为 XXX 的大客车沿新源路由北向南行驶至曹安公路路口处遇绿灯，遂右转弯由东向西行驶，适逢被害人李红英骑电动自行车沿新源路西侧非机动车道由北向南行驶至此，两车相撞。

① S–P–E 空间三要素标注规范：https://2030nlp.github.io/Sp22AnnoOL/task2_guide.html。

表2 文本空间语义角色标注示例

实体S	空间信息P				事件E	时间T
	处所	路径	方向	终点		
大客车	—	沿新源路	由北向南	至曹安公路路口处	行驶	2020年7月16日7时8分许
大客车	—	—	由东向西右转弯	—	行驶	遇绿灯之后
电动自行车	—	沿新源路西侧非机动车道	由北向南	至曹安公路路口处	行驶	2020年7月16日7时8分许
两车	曹安公路路口处	—	—	—	相撞	—

跟上文提到的S–P–E三要素标注相比，表2展示的空间语义角色标注多了一个时间（T）要素。我们把这个语义角色标注体系称为“STEP空间语义角色标注体系”。其中P细分为10个空间角色（如“处所、方向、朝向、起点、终点、路径……”等等），对每个“事件E”，还要进一步标注论元角色“施事”“受事”等。此外，语料中“此”指“曹安公路路口”，“两车”指“大客车”和“电动自行车”，这样的同指（co-reference）信息也需要标注。限于篇幅，这里不展开讨论，有兴趣的读者可访问SpaCE2022网站查询STEP空间语义角色标注规范。①

（五）空间表达异形同义判别

空间方位词在实际使用中，存在语义对立消失的现象，比如“汽车上有炸弹=汽车里有炸弹”。这也正是上文说过的，方位词属于指示语范畴，其具体的空间方位所指，需要更多的认知加工参与，其形式和意义之间的对应关系比其他实词类表达更为复杂。下面来多看几个这种“异形同义”的例子，即句子间存在空间表达的形式差异（通常是一词之差），但不同形式却可以指相同的空间场景。例如：

（5）a. 至今菲律宾的土著居民在见面时，握过手后还要转身**向后**走几步，意思是向对方表明背后没有藏刀。

b.……握过手后还要转身**向前**走几步，……

（6）a. 昨晚，**饭桌上**，奶奶、爸爸和我争着同妈妈说话，直到我双手将妈妈的脸扳向我为止。

b. 昨晚，**饭桌旁**，……

（7）a. 在一座小县城的一间教室里，工人们正在安装一块电子白板。“借助网课，我们的学生坐在教室里，就可以跟着**里面**的名师学习，享受优质的教育资源。”校长兴奋地说。

b.……就可以跟着**外面**的名师学习，……

c.……就可以跟着**上面**的名师学习，……

例（5）两个句子一句是“向后”，一句是“向前”，形式有别，但整句所表达的空间场景信息实际上并无区别。（5a）的“向后”是相对于“转身之前”的方向而言，（5b）的“向前”则是相对于“走”的方向而言，即（5a）是“转身向后”，（5b）是“向前走”。表面上（5a）和（5b）在相同的位置上“后”跟“前”对立，但这个形式上的差异仅仅是表层线性字符串层面的差异，从语言学深层句法结构的层面来看，（5a）跟（5b）是相同的结构，即“转身+走几步”，表层的“向后”或“向前”可以

① STEP标注规范文档网址：https://2030nlp.github.io/Sp22AnnoOL/task3_guide.html。值得一提的是，STEP语义角色标注体系比已有的面向英文的空间语义角色标注数据集如SpRL2012，SpRL2013，SpaceEval2015（参见Kordjamshidi et al. 2012；Kolomiyets et al. 2013；Pustejovsky et al. 2015）要丰富，而且还标注了跟空间语义理解有关的同指关系、时间信息等。

删去而不影响句子的语义。

例（6）“饭桌上”跟“饭桌旁”的对立消失，类似于“汽车上”有时候相当于“汽车里”，“大门前”有时候相当于“大门外”，都跟方位词的多义性，以及空间认知图式有关。

例（7）中在相同位置有 3 个方位词“里面、外面、上面”形成形式上的对立差异，但由于方位词前可补回的参照成分不同，实际上整句可以表达完全相同的场景。

空间表达异形同义包含了不同的类型。除方位词的对立有时会消失外，趋向动词也有类似现象。汉语语法学界讨论较多的所谓“主宾换位”现象，有的也属于空间表达异形同义。如“门口站两个人”“两个人站门口”，“北大西门正对着蔚秀园东门”“蔚秀园东门正对着北大西门”，等等（参见第 4 节表 11）。有关空间表达异形同义现象，我们拟专文讨论，这里不展开。

（六）空间方位关系推理

要考察机器对文本空间信息的综合理解水平，最合适的任务是空间方位关系推理。下面是一个空间关系推理题例子。

（8）桌上有三块积木，红的在绿的上面，黄的在绿的下面。现在把最下面的拿到最上面来。移动之后，中间的积木是什么颜色的？

例（8）中包含 3 个实体，题面涉及“上面、下面”，以及隐含的“中间”方位（题面上未出现，仅在题干部分出现）。像这类涉及实体较少、空间关系相对单一的推理任务，大语言模型的表现比较好。[①] 但实际上，当空间中实体数量变多、空间关系类型增加之后，大语言模型在空间关系推理任务上的表现会出现显著下降。下文将介绍我们在 SpaCE2024 中制作空间关系推理题的情况（见表 8），以及大语言模型在这一任务上的表现（见表 9）。

三、空间语言理解能力评测数据集制作

上一节提出了针对文本空间语义理解的 6 项任务。要将这些任务转化落实为对机器空间语言理解能力的考试，就需要制作一定规模的数据集。表 3 概括呈现了 6 项任务的类型及数据集制作方式，之后对 SpaCE 系列评测的数据制作的总体情况做简要讨论。

表 3　空间语义理解能力测试任务类型及数据制作方式

项目	语言现象	评测任务类型	数据集（试题）制作方法
任务 1	空间信息正误判别	二分类	词语替换 + 人工标注
任务 2	异常空间信息识别	信息抽取	机器辅助人工标注
任务 3	缺失参照成分补回	实体识别（依存分析）	机器辅助人工标注
任务 4	空间语义角色标注	序列标注 / 信息抽取	机器辅助人工标注
任务 5	空间表达异形同义判别	二分类 + 解释理由	机器生成 + 人工编写
任务 6	空间方位关系推理	推理	程序自动生成

为了制作包含异常空间信息的语料，我们首先对自然语料文本中跟空间方位信息有关的词语进行替换，然后由人工来判断替换后的文本是否存在空间信息异常，同时对文本中的异常信息片段进行标

① Kosinski（2023）以物品收纳（20 题）和物品转移（20 题）两类经典的儿童心智能力测试任务来考察 GPT–4，分别得到 90 分和 100 分，以至于 Kosinski 认为 GPT–4 具备了 7 岁儿童的心智能力。但不久之后 Ullman（2023）就设计实验对 Kosinski 的结论提出了质疑。

注和分类，这样就可以得到任务 1 和任务 2 的数据（詹卫东，等 2022）。其中空间标记词 390 词，空间实体词 632 词。[①] 表 4 中的空间标记词就是候选的替换词。根据词语的实际用法特点，分布相近的词语构成替换词族，由程序扫描原始语料，将一段语料中的空间标记词批量替换成同一个替换词族中的其他词语，形成新的语料，进入标注流程。

表 4　SpaCE2024 空间标记词表

空间标记类型	词性标记类型	词语数量	例词
定位标记	方位词	236	上、上面、上边、上头、上端、左、右
	动词	7	在、有、是、位于、居于
	副词	3	到处、处处、四处
	介词	2	在、于
行程标记	动词	22	出发、启程、离开、撤退、抵达
	介词	11	到、至、从、自、由
趋向标记	动词	22	进、出、上、下、回
形态标记	形容词	19	高、低、矮、宽、窄
	动词	10	直、弯、屈、倾斜、弯曲
方向标记	动词	13	进、退、升、降、前进
	副词	6	正向、逆向、同向、同方向、反向
	介词	3	向、向着、往
朝向标记	动词	11	仰、俯、侧、扭、转
	介词	6	朝、朝着、向、向着、对
	区别词	2	向心、离心
距离标记	动词	11	贴、靠、靠近、接近、毗邻
	形容词	4	远、近、深、浅
	介词	2	距、离
合计		390	

图 2 是 SpaCE2022 数据集制作工作流程图，全面展示了文本空间语义理解能力测试中主要数据集（任务 1 到任务 4）的制作工作步骤（图中标记了 12 个主要环节）。

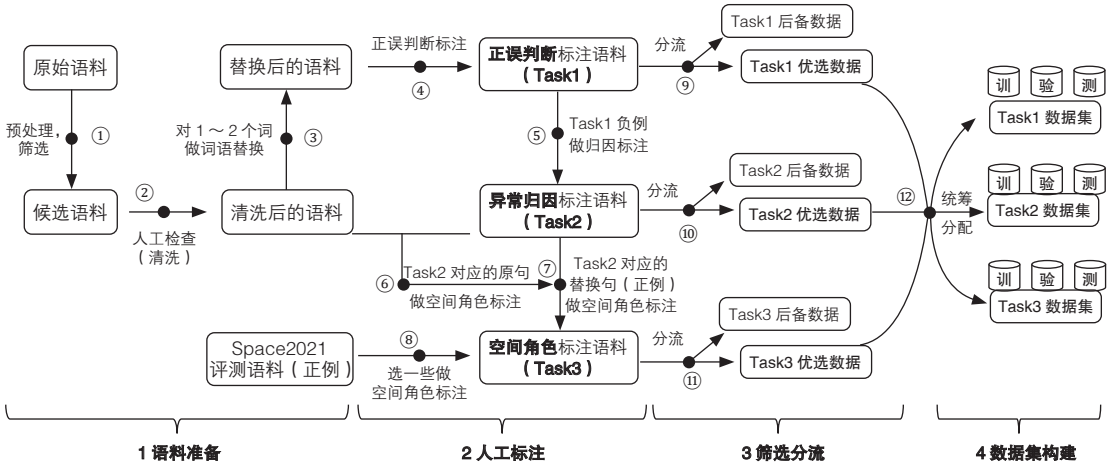

图 2　SpaCE2022 数据集制作工作流程图

① 词表下载网址：https://github.com/2030NLP/SpaCE2024/tree/main/data。

图 2 中，标注 ⑫ 位置的工作是对数据集中的语料进行分配。除一般机器学习常见的划分训练、验证、测试数据集外，在 SpaCE 系列测试任务中，还要在不同子任务中放置一定比例的同源题。如上文例（7），同样的题面，既可以用作异形同义判别的任务，也可以用于缺失参照实体补回的任务，因为这个例子中的 3 个替换词语位置分别是“里面”“外面”“上面”，3 句话属于异形同义句组，而造成异形同义的原因，正是这 3 个方位词前面省略的参照物实体不同：（7a）是“（网课）里面”，（7b）是“（县城）外面”，（7c）是“（电子白板）上面”。表 5 统计了 SpaCE2022 和 SpaCE2023 两届评测中同源题的占比，SpaCE2022 的同源题 ① 在测试集中占比不足 15%，这对分析机器在不同任务上表现的相关性是不利的。因此，SpaCE2023 将同源题 ② 在测试集中的占比提高到了超过 45%（表 10 报告了机器在 SpaCE2023 同源题上的成绩相关性）。

表 5　SpaCE2022 和 SpaCE2023 评测任务中同源题数据统计表

评测	数据集	总题量	原句数	替换句数	同源句数	同源题数	同源题占比 / %
SpaCE2022	task2	7068	0	7068	1351	3476	49.18
	task3	2132	1994	128	1351	1351	63.37
	task2–test	1402	0	1402	59	66	4.71
	task3–test	396	387	9	59	59	14.90
SpaCE2023	task1	7047	0	7047	1681	4562	64.74
	task2	2163	2008	155	1681	1681	77.72
	task1–test	1385	0	1385	220	632	45.63
	task2–test	427	401	26	220	220	51.52

任务 5 很难从自然文本中收集语料：采用替换词语方法造成的“异形”句对，很难恰好满足“同义”条件。这一任务的数据制作主要针对特定的方位词对、趋向动词对，依靠人工编写来得到符合“异形同义”条件的语料，同时借助大语言模型的生成能力来生成候选语料。表 6 展示了 GPT–4 辅助生成“异形同义”句对语料的例子。

表 6　利用大语言模型生成“异形同义”句对示例

提示语（prompt）	请写一段话，其中包含人或物的空间场景信息。这段话中必须包含“前”这个词。然后把这段话拷贝一份，再将拷贝后的文本中的“前”替换成“下”。要求前后这两段文本中涉及的人、物对应的空间场景相同。也就是说，尽管两段话中有一处词语差异，“前”和“下”是不同的，但整段话所描述的空间场景却是相同的。
GPT–4	原文：在这个宽敞的图书馆中，一位年轻女孩安静地坐在书架前，认真阅读着一本书。 拷贝文本：在这个宽敞的图书馆中，一位年轻女孩安静地坐在书架下，认真阅读着一本书。
提示语（prompt）	请写一段话，其中包含人或物的空间场景信息。这段话中必须包含“走到最里面”这个词。然后把这段话拷贝一份，再将拷贝后的文本中的“走到最里面”替换成“走到最前面”。要求前后这两段文本中涉及的人、物对应的空间场景相同。也就是说，尽管两段话中有一处词语差异即“走到最里面”和“走到最前面”是不同的，但整段话所描述的空间场景却是相同的。
GPT–4	原文：在这个宽敞的超市里，顾客们可以在琳琅满目的商品中挑选自己喜欢的东西，当他们走到最里面，便能发现一个美食区，提供各种美味的熟食和甜点。 拷贝文本：在这个宽敞的超市里，顾客们可以在琳琅满目的商品中挑选自己喜欢的东西，当他们走到最前面，便能发现一个美食区，提供各种美味的熟食和甜点。

① SpaCE2022 的 task2 是“异常空间信息识别”任务，task3 是“空间语义角色标注”任务。

② SpaCE2023 的 task1 是“异常空间信息识别”任务，task2 是“空间语义角色标注”任务。

任务 6 同样也难以从自然语料中获取，人工编写推理题成本太高，而且很难保证试题质量。为此，我们提出了一种基于空间布局知识库由程序自动生成推理题的数据合成（data synthesis）方法（将另文讨论）。下面是一个示例。

（9）猪八戒、高翠兰、东海龙王、铁扇公主四人来到茶餐厅吃饭，坐在四人卡座上。卡座分东西两排，每排坐两人，坐东边的两人面朝西，坐在西边的两人面朝东，两排人面对面而坐。已知：高翠兰面朝东且在东海龙王左手边挨着坐，猪八戒右手边坐着铁扇公主。请问：高翠兰坐在 _____ 正对面。（图 3）

图 3　四人卡座推理题示意图

程序自动出题方法的基本思想是：基于一个已知的空间布局（其中实体数量确定、实体方位关系固定），由程序从该空间布局的知识库文件（包含该空间布局的全部实体方位信息的陈述和推导规则）中随机抽取（生成）n 条命题，该 n 条命题须能够还原出一个完整的空间布局，然后以其中 n−1 条命题构成题面，剩下的 1 条命题中去除 1 个空间实体或空间关系词，构成问题，即生成 1 道有效的推理题。目前已实现四人卡座，六人向心（面对）围坐，六人离心（背对）围坐，三层两列置物架等 4 类空间布局，生成了近 4000 道推理题[①]。

我们按照上述数据制作方法完成了从 SpaCE2021 到 SpaCE2024 的数据加工，语料规模如表 7 所示。

表 7　SpaCE 系列评测任务数据集（语料及最终试题）规模统计表

统计项	SpaCE2021	SpaCE2022	SpaCE2023	SpaCE2024
初始语料池字数	42 000 000	89 740 377	—	—
标注原句句数	463	6643	—	—
标注原句字数	45 574	698 476	—	—
数据集原句句数	402	5095	—	6868
数据集原句字数	40 091	543 316	—	1 009 658
标注语料不重复句数	16 692	47 920	—	—
标注语料不重复句字数	2 088 037	5 557 501	—	—
数据集不重复句数	7738	18 907	—	7654
数据集不重复句字数	863 669	2 158 803	—	1 115 583
数据集题数	**18 236**	**24 947**	**9565**	**10 373**
数据集总字数	2 034 101	2 849 795	1 079 826	1 859 261

需要说明的是，SpaCE2023 和 SpaCE2024 语料均取自 SpaCE2022 语料池（近 0.9 亿字）。该语料池中一般性语料占比 83%，专业领域语料占比 17%。前者包括：报刊语料（36%），文学作品语料（25%），中小学语文课本语料（20%），语言学空间研究相关论文例句语料等其他类语料（2%）；后者包括：交通事故判决书语料（9%），体育动作训练教材语料（6%），地理百科语料（2%）。最终得到 SpaCE2022 数据集总共约 2.5 万条语料，每条语料长度范围为 16 ～ 256 字，平均长度 114.23 字，标准差 49.64 字，总字数约 285 万。

① 目前生成的题量比英文的同类数据集 SpartQA（Mirzaee et al. 2021）少，后者为 138 857 题。但 SpaCE2024 空间关系推理题中包含的空间布局类型、实体间空间方位关系的种类（32 种）远多于 SpartQA（7 种）。这意味着更高的推理难度。

SpaCE2023 对已有数据做了质量优化处理，未增加新的语料；SpaCE2024 在原有基础上小幅扩充了语料标注规模，同时用程序生成了一部分语料（空间关系推理题）。SpaCE2024 对全部任务统一采用选择题形式来命题，覆盖了 5 个任务：异常空间信息识别（DSA），缺失参照实体补回（RSR），空间语义角色标注（ISR），空间表达异形同义判别（RSE），空间方位关系推理（SPR）。表 8 是 SpaCE2024 数据集各类任务数据规模统计表。

表 8　SpaCE2024 数据集详细信息统计表

子任务	训练集		验证集		测试集		合计	
	单选	多选	单选	多选	单选	多选	单选	多选
DSA	1077	0	40	0	530	0	1647	0
							1647	
RSR	937	161	226	24	513	87	1676	272
							1948	
ISR	1074	19	186	4	776	24	2036	47
							2083	
RSE	4	1	44	11	541	139	589	151
							740	
SPR	909	301	468	207	1533	537	2910	1045
							3955	
合计	4001	482	964	246	3893	787	8858	1515
	4483		**1210**		**4680**		**10 373**	

四、机器空间语义理解能力评测结果简要分析

关于 SpaCE 历届赛事中参赛系统的表现，可参考詹卫东等（2022）、肖力铭等（2023a，2023b），也可在 SpaCE 评测网站查询详情。本节对机器表现值得关注的几个方面略做简要分析。表 9 呈现了 SpaCE2024 赛事部分参赛系统在 5 项任务上的得分情况。

表 9　SpaCE2024 部分大语言模型成绩单（总分前 3 名 + 2 个基线系统[①]）

系统	总分	单选题	多选题	ISR	RSR	DSA	RSE	SPR
第 1 名	60.24	64.90	37.45	93.64	89.47	84.80	56.31	34.71
第 2 名	59.69	64.34	36.93	91.43	84.91	81.00	54.31	37.16
第 3 名	59.49	63.55	39.66	94.29	77.19	78.00	58.77	37.11
基线 1	47.92	54.37	16.38	88.18	75.09	68.60	42.00	21.96
基线 2	46.29	52.94	13.78	85.19	54.39	65.60	45.08	25.00

SpaCE2024 的参赛系统全部采用大语言模型完成任务（具体选择的模型及采取策略各有不同）。表 9 按照系统在 5 项任务上表现优劣从左到右排序。可以看到，模型在空间语义角色标注、缺失参照

① 基线 1 是开源系统：Qwen1.5-7B-Chat 微调后的模型（10 亿参数级别）；基线 2 是闭源系统 GPT-4-1106-preview 模型（千亿或万亿参数级别）。总分第 1 名的系统仅采用 Qwen1.5-7B-Chat 开源模型微调完成答题，第 2 名系统除 Qwen1.5-7B 外还采用其他模型投票选出答案，第 3 名系统采用 ChatGPT-4o 闭源系统多次投票选出答案。在空间关系推理题上，总分第 1 名的系统成绩排第 3。

成分补回、异常空间信息识别等任务上表现更为出色，而在异形同义判别、空间关系推理任务上表现不佳，尤其是空间关系推理任务，最好成绩也不到 38 分。值得注意的是，在语义角色标注、缺失参照成分找回、异常空间信息识别任务上，基线 1 的成绩超过基线 2，说明对于常见的任务类型，微调效果显著；而对于异形同义判别、空间关系推理等难度更高、训练样本数据较少的任务类型，基线 1 表现低于基线 2，微调没有明显效果，模型的参数规模起到更显著的作用。

下面再简要看一下 SpaCE2023 的 3 个子任务上的机器表现情况。表 10 呈现了基线系统（基于 BERT 微调模型）在 SpaCE2023 的 task1（异常空间信息识别）和 task2（空间语义角色标注）两个子任务同源题上得分的相关性。

表 10　基线模型（BERT）在 SpaCE2023–task1 和 task2 同源题上得分相关系数（P = 0.01）

统计项		基线模型
基于 task1 同源题 (632 题) 看 t1–t2 得分的相关系数		0.10
基于 task2 同源题 (220 题) 看 t2–t1 得分的相关系数		0.17
对齐标注——task1 与 task2 得分的相关系数		0.10
对齐角色——task1 与 task2 得分的相关系数	S	0.12
	P	0.08
	E	0.17

BERT 微调模型在 task1 和 task2 两个任务上的总体表现不佳，得分分别是 0.55 和 0.48，模型在两个任务同源题上的成绩相关系数非常低。这说明机器在完成空间语义理解任务时，即便是同一领域的不同子任务，仍然可能是针对特定任务形式进行学习，而没有“打通”底层的语义逻辑。

表 11 呈现了大语言模型在 SpaCE2023–task3（异形同义判别任务）上的表现。因试题数据规模不大（只有 100 题），不一定能从大语言模型的表现得出可靠的结论。但值得注意的是，ChatGPT–3.5 在**异形同义**题上的得分比**异形异义**题低 12 个百分点，应该能反映大语言模型的“知识能力”仍具有非常突出的“数据驱动”特点，因为在日常语料中，绝大多数的情况都是“异形异义”，这属于常规数据。相比之下，“异形同义”是从语言学研究者角度特别关注的稀有超常规数据，大语言模型的表现相对较差，也就在情理之中了（可对照 Liu et al.（2023）基于“分布外（out of distribution，OOD）数据”对大语言模型推理能力的测试研究[①]）。

表 11　ChatGPT–3.5 在 SpaCE2023–task3（100 道测试题）上的表现

异形同义类别	总题数	同义题数	异义题数	判断	解释	同义题分	异义题分	总分
方位图式相同	57	24	33	35(61.40%)	22.4(39.30%)	0.36	0.44	0.39
趋向动词	24	13	11	20(83.33%)	10.0(41.67%)	0.32	0.53	0.42
方位词义包含	6	6	0	6(100.00%)	2.8(46.67%)	0.47	—	0.47
参照物补回	6	4	2	5(83.33%)	2.2(36.67%)	0.43	0.65	0.37
主宾逆序	4	4	0	3(75.00%)	1.7(42.50%)	0.23	—	0.43
方位词义相同	2	2	0	1(50.00%)	0.7(35.00%)	0.35	—	0.35
实体投影关系	1	1	0	1(100.00%)	0.0(0.00%)	0.00	—	0.00
合计	100	54	46	71(71.00%)	40.4(40.40%)	0.35	0.47	**0.40**

① Liu et al.（2023）构建了 AR-LSAT 数据集，评估 GPT–4 的逻辑推理能力。GPT–4 在 AR-LSAT 数据集上的成绩是 33.48 分（满分 100 分）。

空间表达的异形同义现象可以从不同角度认识和分类。对此我们将另文讨论。这里简要说明表 11 中的分类：方位图式相同指“上–里”同义的情形；趋向动词指“插上–插下”同义的情形；方位词义包含指“上端–顶端”类同义情形；方位词义相同指“里–中–内”同义情形；实体投影关系指“镜头前–镜头里”同义的情形。SpaCE2023–task3 要求机器分两步来答题，先判断一组句对属于“异形同义”句还是“异形异义”句，然后再按照一定的模板格式，解释判断理由。表 11 中分“判断”和“解释”两列呈现了机器表现情况。①

总的来看，对大语言模型来说，SpaCE 系列空间语义理解能力评测，依然是高挑战性任务。解决问题的线索越是依赖表面分布特征（形式线索），机器就越容易获得好的成绩，而越是依赖深层语义理解的任务（认知能力），或者可获得的训练样本数据量越小的任务，机器就越容易表现不好。本文没有介绍 SpaCE 系列任务上人类测试的情况（可参考詹卫东，等 2022）。值得注意的是，对于空间语义理解中“异形同义判别”这类凸显认知加工主观性的任务，初步的人类测试结果也显示了不同个体之间较为突出的不一致性，对此将另文讨论。

五、结　语

本文较为全面地介绍了基于语言学理念设计 SpaCE 系列评测任务以及相应的数据集制作工作的总体情况。这项研究还有许多可改进之处，比如测试题对各类空间语言现象的覆盖率，试题内部的结构化设计，包括难度在内的更多更灵活的特征变量控制，等等。

在面向大语言模型的语言能力测试研究工作中，本文引言中提出的问题——知识和数据的关系——具体化为：如何依据语言知识提出好的语言能力测试问题，制作出高质量的测试题（数据）。如果语言知识真的足够可靠，就可以基于语言知识，由程序来自动生成数据。SpaCE2024 中的空间关系推理题，就是从人类语言知识出发实现由程序自动出题的一次尝试。更进一步，我们需要思考：能否在更大范围内、更系统地进行类似的实践？换言之，如何让语言学知识来系统地指导“从知识生成数据”的语言工程实践？在符号主义 AI 时代，语言学知识的价值在于以程序可读知识库形式，直接用“显性符号知识”去武装机器的“大脑”；在当前的联结主义 AI 时代，语言学知识的价值需要重新定位，即用于指导生产小而精的高品质语言数据，人类不再直接以知识，而是以数据（即语料）形式来“喂养机器的神经网络”，实现提升机器学习的效果和效率的目标。

乔姆斯基开创的“**生成**语言学”革命，首次把语法视作一种“生成装置”，即“一个语言（L）的语法将是一个生成所有 L 序列而不生成任何非 L 序列的装置”（Chomsky 1957，第 2 章）。如果语法真的能做到这一点，也就实现了乔姆斯基为语法研究勾画的“All and only”的宏伟蓝图（Chomsky 1957，第 8 章）。在今天的时代背景下，为了计算应用的目的，或许在乔姆斯基当年明确提出的语法研究的 3 个目标“对语言现象的观察充分（observational adequacy）、描写充分（descriptive adequacy）、解释充分（explanatory adequacy）”（Chomsky 1965，第 1 章）之上，还应该升格一个更具挑战性的目标——**生成充分（generative adequacy）**。

1965 年，诺贝尔物理学奖得主、物理家理查德·费曼（Richard Feynman）在黑板上留下一句名

① SpaCE2023–task3 评分标准可参见 https://2030nlp.github.io/SpaCE2023/#eval。

言：“What I cannot create, I do not understand.”（一个事物可理解的前提是我能创造它。）[①] 创造了 ChatGPT 的 OpenAI 公司同样把这句话作为生成式人工智能的宣言。[②] 无独有偶，世界著名华人数学家丘成桐先生也表达过类似的看法：“理解几何结构最透彻的方法就是弄明白如何从零开始构建几何结构。”[③] 这些思想，跟生成语法学的理论追求一致，或许都可以概括为：检验知识的最佳手段，就是用知识去生成数据。

语言学家的大脑应该是比计算机程序更厉害的语言数据生成器。比如乔姆斯基创造的那些例句，实际上对于今天的大语言模型来说，仍然构成挑战。[④] 在深度学习技术为主流的 AI 研究中，通过知识生成数据，再将数据用于模型训练和测试，不仅可以检验知识的可靠性，更可以直接助力提升模型的性能。而能够用知识生成数据的前提，正是“知识是可靠的”。反过来，如果语言学知识无法生成出“正确的语言数据”，人们就有理由怀疑：已有的语言学知识不可靠，或者还不够可靠。因此，语言学工作者应该借助计算机程序，或者借助自己的脑力，将更多语言学研究成果转化为语言数据。知识是不是有效，生成数据测一测才知道。在 AI 时代，语言学理论的价值是用理论生成（创作）例句，而不仅仅是用理论解释例句。

参考文献

董青秀，穗志方，詹卫东，等　2021 《自然语言处理评测中的问题与对策》，《中文信息学报》2021 年第 6 期。

肖力铭，孙春晖，詹卫东，等　2023a 《SpaCE2022 中文空间语义理解评测任务数据集分析报告》，载《第二十二届中国计算语言学大会论文集》，https://aclanthology.org/2023.ccl-1.48/。

肖力铭，詹卫东，穗志方，等　2023b 《CCL23–Eval 任务 4 总结报告：第三届中文空间语义理解评测》，载《第二十二届中国计算语言学大会论文集》，https://aclanthology.org/2023.ccl-3.14/。

詹卫东　2024 《如何评估机器的语言能力？》，载杨旭、罗仁地主编《ChatGPT 来了——语言科学如何看待 ChatGPT》，上海：上海教育出版社。

詹卫东，孙春晖，岳朋雪，等　2022 《空间语义理解能力评测任务设计的新思路——SpaCE2021 数据集的研制》，《语言文字应用》2022 年第 2 期。

Chollet, F. 2019. On the Measure of Intelligence. https://arxiv.org/abs/1911.01547.

Chomsky, N. 1957. *Syntactic Structure*. The Hague: Mouton.

Chomsky, N. 1965. *Aspects of the Theory of Syntax*. Cambridge Mass.: MIT Press.

Chomsky, N. 1986. *Knowledge of Language: It's Nature, Origin and Use*. New York: Praeger Publishers.

Chomsky, N., I. Roberts & J. Watumull. 2023. The false promise of ChatGPT. *The New York Times*, March 8. https://www.nytimes.com/2023/03/08/opinion/noam-chomsky-chatgpt-ai.html.

Hinton, G. 2024. 辛顿教授在爱尔兰大学获颁尤利西斯奖章（UCD Ulysses Medal）演讲致辞. https://www.youtube.

① https://digital.archives.caltech.edu/collections/Images/1.10-29/。

② https://openai.com/index/generative-models/。

③ 丘成桐《数学的万有引力》，载微信公众号“数理人文”2023 年 11 月 14 日。https://qzc.tsinghua.edu.cn/info/1017/4641.htm。

④ 笔者用下面的例子测试，包括目前最好的 GPT–4o 大模型在内，也并不能做出完全正确的分析。

(1) a. The mechanic who fixed the car carefully packed his tools.

b. Carefully, the mechanic who fixed the car packed his tools. 两句意思有区别吗？

(2) John is too stubborn to talk to. 这句话是什么意思？“talk to”应理解为谁跟谁说话？

(3) There (　) a pen and two books on the desk. 括号内填什么单词？

详见 https://github.com/d0ubtfire/LLM_Evaulation/tree/main/ 对比大模型 /Linguistic Knowledge and Proficiency Test。

com/watch?v=IIl2DbLvBtE. 又见《杰弗里·辛顿接受尤利西斯奖章时发表的获奖感言》，陈国华，译，《当代语言学》2024 年第 4 期。

Katzir, R. 2023. Why large language models are poor theories of human linguistic cognition: A reply to Piantadosi. Biolinguistics, 17, Article e13153. https://doi.org/10.5964/bioling.13153, https://lingbuzz.net/lingbuzz/007190.

Kolomiyets, O., P. Kordjamshidi, S. Bethard, et al. 2013. Semeval–2013 task 3: Spatial role labeling. *Proceedings of the Seventh International Workshop on Semantic Evaluation (SemEval 2013)*. 255–262.

Kordjamshidi, P., S. Bethard & M. F. Moens. 2012. SemEval–2012 task 3: Spatial role labeling. *Proceedings of the Sixth International Workshop on Semantic Evaluation (SemEval 2012)*. 2, 365–373.

Kosinski, M. 2023. Theory of mind may have spontaneously emerged in large language models. Stanford University, https://arxiv.org/abs/2302.02083.

Legg, S. & M. Hutter. 2007. Universal intelligence: A definition of machine intelligence. *Minds & Machines* 17(4), 391–444.

Levesque, H., E. Davis & L. Morgenstern. 2012. The Winograd Schema Challenge. *Thirteenth International Conference on the Principles of Knowledge Representation and Reasoning*, 552–561.

Liu, H., R. Ning, Z. Teng, et al. 2023. Evaluating the logical reasoning ability of ChatGPT and GPT–4. https://arxiv.org/abs/2304.03439.

Mirzaee, R., H. R. Faghihi, Q. Ning, et al. 2021. SPARTQA: A textual question answering benchmark for spatial reasoning. *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies.* 4582–4598.

Piantadosi, S. 2023. Modern language models refute Chomsky's approach to language. https://lingbuzz.net/lingbuzz/007180.

Pustejovsky, J., P. Kordjamshidi, M. F. Moens, et al. 2015. Semeval–2015 task 8: Spaceeval. Proceedings of the 9th International *Workshop on Semantic Evaluation*. 2015, 884–894.

Sakaguchi, K., R. L. Bras, C. Bhagavatula, et al. 2020. WinoGrande: An adversarial Winograd schema challenge at scale. *Proceedings of the AAAI Conference on Artificial Intelligence* 34(5), 8732–8740.

Ullman, T. 2023. Large language models fail on trivial alterations to theory-of-mind tasks. Harvard University, https://arxiv.org/abs/2302.08399.

Wolfram, S. 2023. What is ChatGPT doing and why does it work? https://writings.stephenwolfram.com/2023/02/what-is-chatgpt-doing-and-why-does-it-work/.

Zhong, W., R. Cui, Y. Guo, et al. 2023. AGIEval: A human-centric benchmark for evaluating foundation models. https://arxiv.org/abs/2304.06364.

责任编辑：王　飙